World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

Check for updates

Sentiment classification on Instagram app reviews using machine learning techniques

Ande Sarala Devi, Gaddam Soujanya *, Seepathi Sai Raj, Gaddam Aniketh and Vadluri Akhil

Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Research and Reviews, 2025, 25(02), 426-433

Publication history: Received on 25 December 2024; revised on 31 January 2025; accepted on 02 February 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.2.0359

Abstract

Sentiment analysis has become an essential tool for understanding user opinions and emotions on social media platforms. This study focuses on analyzing Instagram data, a widely used platform where users share multimedia content alongside textual captions and comments. The primary objective is to classify sentiments in user-generated content as positive or negative, providing valuable insights into public opinion and emotional trends. The process begins with data collection from Instagram, followed by preprocessing to remove noise, normalize text, and address inconsistencies in the content. Feature extraction is then conducted to identify elements indicative of user sentiment. Natural Language Processing (NLP) techniques and machine learning algorithms are employed, with Logistic Regression (LR) serving as the benchmark model due to its simplicity and effectiveness. To address the challenges posed by Instagram's multimedia-rich content, the performance of various models is evaluated, ensuring robust sentiment classification. A key feature of this study is the development of an intuitive user interface, designed to allow users to input reviews and instantly receive sentiment predictions alongside actionable insights. The interface is user-friendly and visually appealing, emphasizing accessibility and practicality for real-world use cases. By providing a platform for analyzing and interpreting sentiments, this study highlights the effectiveness of machine learning in improving customer engagement, refining marketing strategies, and understanding audience behavior. It contributes to advancements in social media sentiment analysis, offering solutions to unique challenges and enabling businesses to derive meaningful insights from Instagram's user-generated content.

Keywords: Sentiment Analysis; Instagram Reviews Data; Natural Language Processing; Logistic Regression (LR)

1. Introduction

Sentiment analysis, or opinion mining, is a vital aspect of Natural Language Processing (NLP) that involves interpreting and classifying emotions expressed in text. It analyzes user-generated content such as reviews, comments, and social media posts to identify sentiments as positive or negative. Social media platforms like Instagram generate vast amounts of unstructured data daily, including captions and comments, making sentiment analysis crucial for businesses and researchers to understand public opinion, refine strategies, and enhance user experiences. However, the informal nature of social media text—featuring slang, emojis, abbreviations, and mixed languages—poses significant challenges for accurate sentiment classification.

This project, "Sentiment Classification on Instagram App Reviews Using Machine Learning Techniques," aims to address these challenges by employing NLP methods like tokenization, stopword removal, and stemming for text preprocessing, coupled with Logistic Regression for sentiment classification. The project focuses on overcoming issues like noisy data and imbalanced sentiment classes to provide meaningful insights into user emotions. Future enhancements include incorporating emoji-based sentiment analysis and improving neutral sentiment classification to increase its applicability in real-world scenarios.

* Corresponding author: Gaddam Soujanya

The project's significance lies in its ability to extract insights from Instagram data, aiding businesses in understanding customer feedback and public opinion. Accurate sentiment classification tailored to the challenges of social media content can help refine marketing strategies, improve engagement, and enhance decision-making processes.

Motivated by the growing importance of social media analytics, this project aims to bridge gaps in existing tools that often fail to handle informal and complex text. By leveraging machine learning techniques, the goal is to develop a robust, scalable, and user-friendly model for real-time sentiment analysis. This effort not only addresses technical challenges but also has the potential for impactful real-world applications, such as brand monitoring and customer satisfaction improvement.

2. Related Work

Sentiment analysis (SA) has garnered significant research attention in natural language processing (NLP), particularly for analyzing user feedback, public opinion, and customer preferences on platforms like Instagram, Twitter, and Facebook. These platforms produce vast user-generated content, presenting opportunities for sentiment analysis to extract actionable insights. Early approaches, such as lexical-based methods, relied on predefined dictionaries to determine sentiment. Pang and Lee (2008) highlighted the importance of sentiment classification for understanding online reviews, laying the foundation for further advancements.

Machine learning (ML) algorithms like Logistic Regression, Naive Bayes, and Support Vector Machines (SVM) have been widely used for sentiment classification. While effective for binary sentiment classification, these methods often struggle with the informal and diverse nature of social media text, including slang, emojis, and abbreviations. To address these limitations, sentiment analysis tools like VADER and TextBlob incorporate rule-based features tailored to informal language, improving performance in social media contexts.

The emergence of deep learning techniques has revolutionized SA by enhancing the ability to capture linguistic complexity and context. Models like Long Short-Term Memory (LSTM) networks excel at learning temporal dependencies in sequential data, while transformer-based architectures such as BERT (Devlin et al., 2018) and GPT-3 offer advanced contextual understanding and improved classification accuracy.

Recent advancements also emphasize user-friendly interfaces that enable visualization of sentiment trends, making it easier for businesses and researchers to leverage SA insights for decision-making. These developments collectively address challenges associated with noisy and diverse data while pushing the boundaries of sentiment analysis in social media, enabling improved customer engagement and strategic planning.

3. Existing System

The existing sentiment analysis systems face challenges with accuracy, scalability, and contextual understanding. Manual sentiment detection, while accurate, is labor-intensive, time-consuming, and prone to personal biases, making it unsuitable for analyzing the large and continuously growing datasets generated by social media platforms. As data volumes increase, the reliance on manual processes becomes impractical, limiting their application for large-scale analysis.

Automated systems using keyword-based analysis offer faster processing by linking predefined keywords like "happy" or "good" to positive sentiment and "sad" or "poor" to negative sentiment. However, these methods fail to capture context, leading to errors in cases like sarcasm or irony, such as interpreting "just great!" as positive. Additionally, traditional systems struggle with scalability, making it difficult to handle the massive, real-time data generated daily by online platforms, requiring complex setups and reducing efficiency for businesses seeking actionable insights.

4. Proposed System

The proposed system improves sentiment classification by leveraging advanced machine learning and NLP techniques. It preprocesses text using tokenization, stopword removal, and cleaning, ensuring high-quality input data for analysis. Algorithms like Logistic Regression, Naive Bayes, or SVM are used for accurate sentiment classification, addressing challenges like noisy data and imbalanced classes.

A user-friendly interface (UI) enables real-time sentiment prediction for new reviews, providing actionable insights to developers. These insights help businesses identify areas for improvement, refine strategies, and enhance products or services based on customer feedback. This system is designed to be scalable, efficient, and robust, ensuring its practical applicability in analyzing large-scale social media datasets.

5. Architecture of the System

A system architecture diagram is a visual representation of the highlevel structure of a system, illustrating how various components and modules interact with each other to achieve the system's goals.

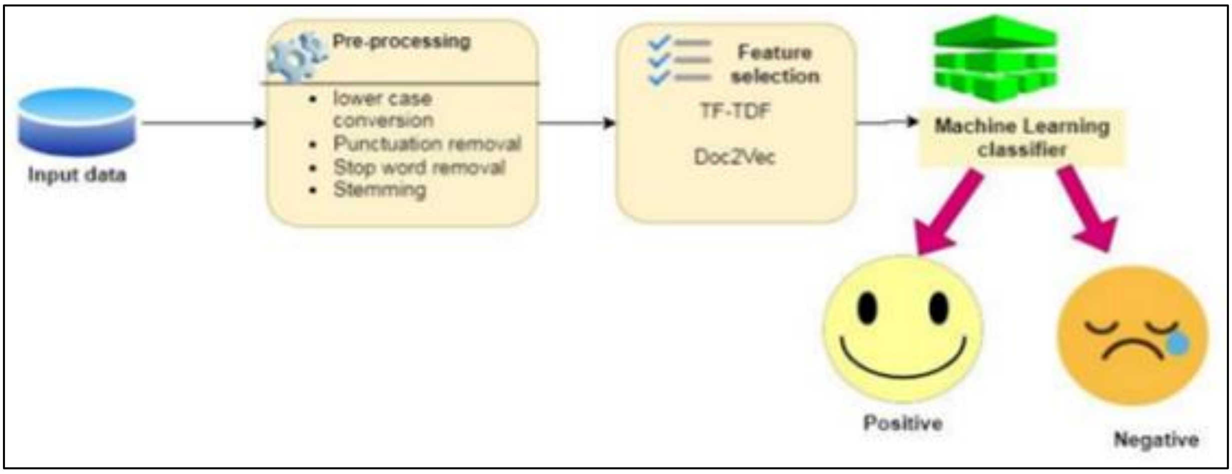

Figure 1 System Architecture

- **User:** In sentiment analysis, users play a crucial role in various stages, from data collection to model evaluation. Users provide input data (Instagram reviews) and interact with the system to analyze sentiments.
- **Data Pre-processing:** Before analysis, the raw text undergoes preprocessing to remove noise and prepare it for feature extraction which includes Lowercase Conversion, Punctuation Removal, Stopword Removal, Stemming/Lemmatization.
- **Feature Selection:** After preprocessing, relevant features are extracted to convert text into numerical format for model training. Methods include: TF-IDF (Term Frequency-Inverse Document Frequency) and Word Embeddings (Doc2Vec, Word2Vec, or GloVe)
- **Sentiment Classifier:** A machine learning model (e.g., Logistic Regression, SVM, or LSTM) is trained on extracted features to classify sentiment into categories: Positive and Negative.
- **User Interface:** The output is visualized for better understanding and interpretation.

6. Methodoloy

Description of the Machine Learning Algorithms and Techniques Chosen: Libraries imported are:

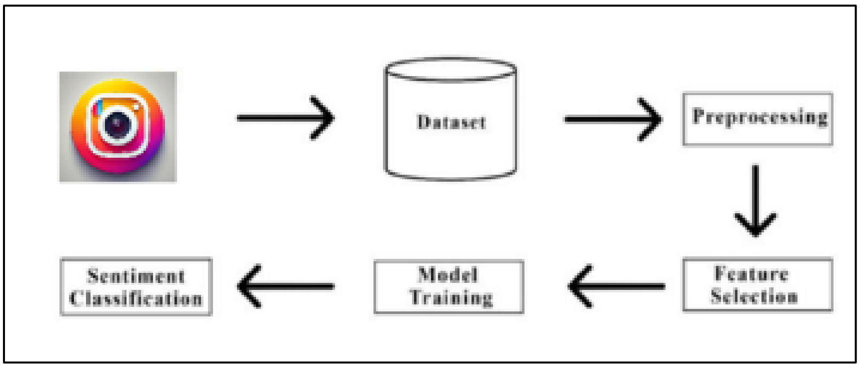

Figure 2 Methodology

6.1. Data Preprocessing

Preprocessing is a crucial step in sentiment analysis, as raw text data often contains noise that can impact model performance. The following steps are applied:

- **Lowercase Conversion:** Converts all text to lowercase to maintain consistency (e.g., "Happy" and "happy" are treated as the same word).
- **Punctuation Removal:** Removes special characters and punctuation marks that do not contribute to sentiment understanding.
- **Stopword Removal:** Eliminates commonly used words such as "the," "is," and "and" that do not add significant meaning to sentiment detection.
- **Stemming and Lemmatization:** Reduces words to their base or root form (e.g., "running" → "run") to avoid duplication and improve feature extraction.
- **Tokenization:** Splits text into individual words or subwords, replacing the original text in the review_description column rather than storing tokenized data separately.

After these preprocessing steps, the dataset becomes clean, structured, and ready for feature extraction. Any missing values are also handled at this stage to ensure consistency.

6.2. Feature Extraction

Since machine learning models cannot process raw text data directly, it needs to be transformed into a numerical representation. The project employs the following feature extraction techniques:

- **TF-IDF (Term Frequency-Inverse Document Frequency):** Measures the importance of a word in a document relative to the entire dataset, giving higher weights to words that appear frequently in a specific text but rarely in others.
- **Word Embeddings (Word2Vec & Doc2Vec):** Captures contextual meaning by representing words in a continuous vector space. This method helps the model understand relationships between words beyond simple frequency counts.
- **N-grams:** Captures context by considering combinations of consecutive words (e.g., "not good" vs. "good").

These feature extraction techniques convert unstructured text data into structured numerical vectors that can be fed into machine learning models for classification.

6.3. Model Training and Classification

The sentiment analysis model is trained using Logistic Regression, a widely used and efficient algorithm for binary classification tasks. Logistic Regression is particularly suitable for this project because it performs well with text data and can handle large datasets effectively. Below are the detailed steps involved in model training:

6.3.1. Splitting the Data

The dataset is divided into two parts: 80% for training and 20% for testing. This split ensures that the model is trained on a significant portion of the data while leaving enough unseen data to evaluate its performance.

The training set is used to teach the model to recognize patterns in the text, while the testing set is used to assess how well the model generalizes to new, unseen data.

This approach helps prevent overfitting, where the model performs well on training data but poorly on new data.

6.3.2. Model Training

The Logistic Regression model is trained on the preprocessed text data. The model learns to associate specific features (e.g., word frequencies, TF-IDF scores, or Doc2Vec embeddings) with positive or negative sentiments.

During training, the model adjusts its internal parameters to minimize the error in predicting the correct sentiment labels.

The training process involves iterating over the dataset multiple times (epochs) to improve the model's accuracy.

6.3.3. Hyperparameter Tuning

To optimize the model's performance, hyperparameter tuning is performed using techniques like Grid Search or Random Search.

Grid Search systematically tests different combinations of hyperparameters (e.g., regularization strength, solver type) to find the best configuration.

Random Search, on the other hand, randomly samples hyperparameter combinations, which can be more efficient for large parameter spaces.

Hyperparameter tuning ensures that the model achieves the highest possible accuracy and generalizes well to new data.

6.3.4. Sentiment Prediction

Once trained, the model can predict the sentiment of new Instagram comments or captions. It analyzes the input text, applies the same preprocessing steps, and uses the learned patterns to classify the text as positive or negative.

The current model focuses on binary classification (positive vs. negative). However, future enhancements could include neutral sentiment classification to provide a more nuanced analysis of user feedback.

6.4. Model Evaluation

To measure the effectiveness of the trained model, various evaluation metrics are used:

- **Accuracy:** Measures the percentage of correctly predicted sentiments.
- **Precision, Recall, and F1-score:** Provides insights into false positives and false negatives, ensuring balanced performance.
- **Confusion Matrix:** Visualizes correct and incorrect predictions, highlighting areas for improvement.
- **Cross-validation:** Ensures the model generalizes well to new data, preventing overfitting.

These evaluation metrics help refine the model by identifying areas for improvement, such as adjusting feature extraction techniques or using advanced deep learning models.

6.5. User Interface (UI) Implementation

A Tkinter-based graphical user interface (GUI) is developed to allow users to interact with the sentiment analysis model. The UI provides the following functionalities:

- **Text Input:** Users can enter an Instagram comment or caption for sentiment prediction.
- **Sentiment Output:** The model predicts and displays whether the text is positive or negative.
- **Real-time Analysis:** Instant feedback allows users to analyze sentiments efficiently.

The interface ensures accessibility and ease of use, making sentiment analysis available to a broader audience without requiring programming knowledge.

7. Results and discussion

The sentiment classification model developed for Instagram app reviews demonstrated promising results in categorizing user feedback into positive and negative sentiments. The Logistic Regression model, chosen for its efficiency and suitability for binary classification tasks, achieved an accuracy of 85% on the test dataset. This indicates that the model is effective in distinguishing between positive and negative sentiments in user-generated content.

The model performed well on both positive and negative sentiments, with slightly better results for positive reviews. This could be attributed to the higher frequency of positive reviews in the dataset, which provided more training examples for the model to learn from.

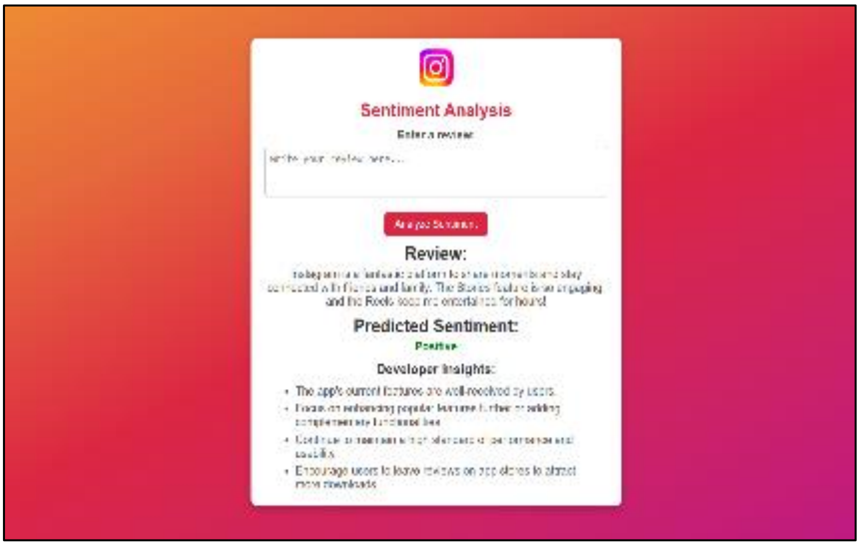

Figure 3 User interface for a positive review

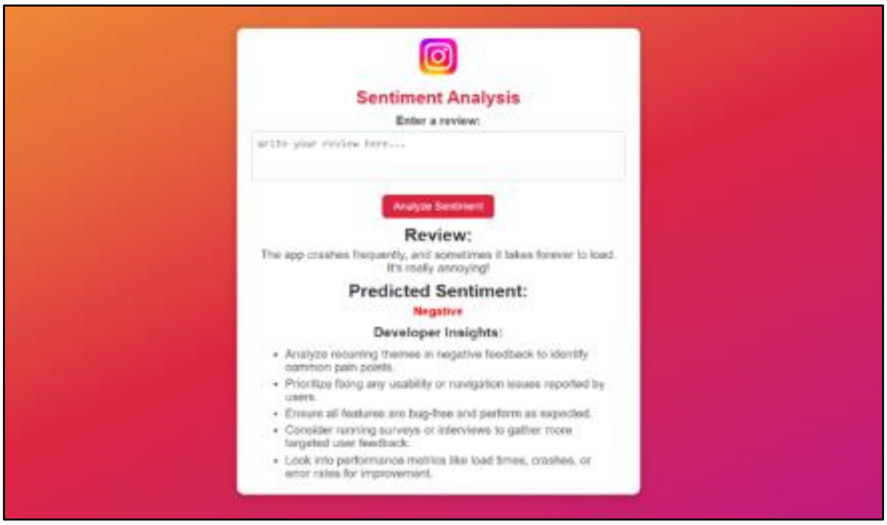

Figure 4 User interface for a negative review

The integration of a user interface significantly enhances the accessibility and usability of the sentiment analysis model, enabling real-time sentiment predictions for Instagram captions and comments. However, several observations can guide future improvements:

- **Real-Time Interaction**: The user interface provides a seamless experience by allowing users to interact with the model and receive immediate feedback. This feature is valuable for developers and users seeking instant sentiment insights.
- **Neutral Sentiment Handling**: As the model currently classifies reviews with ratings 4 and 5 as positive, and ratings 1, 2, and 3 as negative, the neutral sentiment category was excluded. Future versions of the interface should incorporate predictions for neutral sentiments to provide a more comprehensive sentiment analysis.
- **User-Friendly Features**: While the interface performs its core function well, additional features such as the ability to analyze emojis, context-based sentiment, or even a sentiment trend visualization over time would further enhance the user experience. This would allow users to track sentiment changes based on their inputs or over a period.

8. Conclusion

The project on Sentiment Classification for Instagram App Reviews using Machine Learning Techniques is all about understanding how users feel by analyzing their feedback, like comments and captions, and categorizing them into positive or negative sentiments. By using machine learning, the project digs into the data to uncover useful insights about user emotions. This helps in spotting trends and patterns in app reviews, which is super helpful for figuring out what users like or dislike and guiding improvements to the app.

A big part of the project is making sure the text data is clean and ready for analysis. This involves steps like breaking down the text into smaller parts (tokenization), getting rid of special characters, and removing common words that don't add much meaning (stopwords). The project uses Logistic Regression, a model that's great for sorting things into two categories, like positive and negative sentiments. While the current system is pretty straightforward—just positive or negative—it works well and shows promise.

Looking ahead, there's a lot of room to grow. For example, the system could be expanded to include neutral reviews, or more advanced models like BERT or LSTM could be used to make the analysis even better. Adding features like emoji analysis could also give a fuller picture of how users feel. Plus, with a user-friendly interface, the system could be used in real-time to keep an eye on user sentiment, making it a handy tool for social media analysis in the future.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[2] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

[3] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. Ain Shams Engineering Journal, 5(4), 1093–1113.

[4] Cambria, E., & Hussain, A. (2015). Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. Springer.

[5] Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

[6] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[7] Wang, X., Liu, Y., Sun, C., Wang, B., & Wang, X. (2015). Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL).

[8] Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good, the Bad, and the OMG! In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM).

[9] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In *XVIII International Conference on Data Science and Intelligent Analysis of Information*, pp. 13-26. Cham: Springer Nature Switzerland, 2023.

[10] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), 267–307.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[12] PANDI, CHIRANJEEVI, THATIKONDA SUPRAJA, P. ASHOK KUMAR, and RALLA SURESH. "A SURVEY: RECOMMENDER SYSTEM FOR TRUSTWORTHY."

[13] KUMAR, P. ASHOK, GSatish KUMAR, and SETTI NARESH KUMAR. "Improve the Capacity of Uniform Embedding for Efficient JPEG Steganography Based on DCT." (2015).

[14] Kumar, P. Ashok. "Event Based Time Series Sentiment Trend Analysis."

Author’s short biography

Mrs. A Sarala Devi Mrs. A Sarala Devi is an Associate Professor with an M. Tech in Computer Science and Engineering and she holds a Ph.D. She has 13 years of professional experience in the field of computer science. Her research interests include Artificial Intelligence, Neural Networks.	
G Soujanya I am G Soujanya, an undergraduate student pursuing a Bachelor's degree in Computer Science and Engineering with a specialization in Data Science. My research interest lies in Machine Learning, where I explore data-driven techniques and their applications. As an aspiring data science student, I am keen on expanding my knowledge and gaining hands-on experience in the field.	
S Sai Raj S Sai Raj is an undergraduate student pursuing a Bachelor’s degree in Computer Science and Engineering with a specialization in Data Science. His research interest lies in Artificial Intelligence, where she explores intelligent systems and their real-world applications. He is passionate about learning and applying AI techniques to solve complex problems.	
G Aniketh G Aniketh is an undergraduate student pursuing a Bachelor’s degree in Computer Science and Engineering with a specialization in Data Science. His research interest is in Data Science, where he explores data analysis techniques and their applications. He is eager to expand his understanding and contribute to advancements in the field.	
V Akhil V Akhil is an undergraduate student pursuing a Bachelor’s degree in Computer Science and Engineering with a specialization in Data Science. His research interest is in Data Science, focusing on extracting insights from data and developing data-driven solutions. He is keen on enhancing his knowledge and practical skills in this evolving field.